COMPARATIVE STUDY OF MULTIDIMENSIONAL ANALYSIS FOR SINGLE CELL DATA

by
Archana Balan

A thesis submitted to Johns Hopkins University in conformity with the requirements for

the degree of Masters in Science and Engineering

Baltimore, Maryland
May 2020

Abstract

Interpreting interpretable cellular and molecular processes from single cell data is hindered by their high dimensionality. Therefore, dimensionality reduction is an important aspect of single cell data analysis. In addition to providing a manageable set of features in a latent space for interpretation, dimensionality reduction methods further filter the noise present in single cell data and reduce the computational intensity of subsequent analyses. However, such analyses are sensitive to the number of features sought through dimensionality reduction. An important aspect of such analysis is obtaining the optimal dimension, which remains an open question in unsupervised learning. Recent work suggests that multi-resolution models that summarize low dimensional features across dimensions provide more accurate interpretation than similar analyses relying on a single, optimal dimensionality. Therefore, this study analyses the effect of dimensionality on the biological relevance of the latent space that is learned using three prominent dimensionality reduction methods for single cell analysis: CoGAPS, ACTIONet and VAE. We have compared the effect of increasing dimensionality on preserving underlying biological hierarchies, recovering cell type annotations and their consistency across dimensions for a publicly available scRNA-seq PBMC dataset.

Thesis Readers:

Dr.Elana Fertig

Dr.Patrick Cahan

Dr.Robert Scharpf

Table of Contents

List of Figures

Introduction

Single cell gene expression data is typically high dimensional and characterized by a large number of molecular measurements (genes) in comparison to the number of samples (cells). The analysis of such data into interpretable biological processes is often computationally intensive and hindered by its high-dimensionality, and further complicated by technical artifacts such as drop out in the data. Therefore, single cell analysis often relies on representations of the data into a lower dimensional space. Various dimensionality reduction methods are used for learning a latent space representation [1] that captures vital biological information of the original high dimensional data.

Dimensionality reduction is widely used for visualization [2], [3], feature extraction [4], [5] and downstream analysis of single cell data [6]. It includes a wide array of methods ranging from factorization methods of principal component analysis (PCA), independent component analysis (ICA), and non-negative matrix factorization (NMF) [7], embedding algorithms such as uniform manifold approximation and projection (UMAP) [2] and t-distributed stochastic neighbor embedding (t-SNE) [8], and emerging machine learning based methods such as autoencoders. Each of the methods use different approaches to learn the latent space and prove to be useful for different types of applications [9], [7].

Briefly, dimensionality reduction techniques identify a small set of features from the larger set of molecular measurements and/or samples in the single cell data. An important aspect of dimensionality reduction is determining the optimal number of features that

define the dimension (or resolution) for the latent space representation. The latent factors should ideally capture and distinguish disparate sources of biological information and technical artifacts such as batch from the original data. While a lower than optimal dimension might lose out on vital information, a higher dimension often captures redundant features or features associated with technical variation in the data.While efforts are being directed towards developing standard computational methods for determining the optimal dimension [10], [11], a vast majority of current methods use ad-hoc measures which are highly subjective and vary for different datasets [12].

Recent work suggests that different features are uncovered at different dimensions, posing instead that no single dimensionality is able to infer all relevant features from a single dataset [7], [9], [13]. Hence, it is critical to understand the effect of dimensionality on the biological relevance of the latent space learned in dimensionality reduction methods for single cell data analysis. Significant efforts have been directed towards comparing various performance attributes [14] and visualization aspects [15] of some of the methods. While existing studies focus on inter-method comparisons [16], we focus instead on the impact of dimensionality within methods using a multidimensional analysis of the same dataset with three particular methods: a non-negative matrix factorization based method (CoGAPS) [17], a network based approach (ACTIONet) [13] , and a Variational Autoencoder [18] . We focus this analysis on scRNA-seq data of peripheral blood mononuclear cells (PBMCs) [19], as they have an established ground truth of cell types and hierarchy that are ideal for evaluating dimensionality within and between methods. We design this benchmark analysis both to assess the impact of dimensionality of features learned in latent space representations as well as simplified sets of marker

genes that are critical for using these representations for annotations of single cell data.

Methods and Materials

CoGAPS: Bayesian non-negative matrix factorization analysis of single cell data and cell type identification

CoGAPS (Coordinated Gene Activity in Pattern Sets) is a Bayesian non-negative matrix factorization algorithm that decomposes a data matrix into two lower dimensional matrices of non-negative values. The input to CoGAPS, a gene expression matrix $D \in \mathbb{R}^{n\,x\,m}$ (n genes and m cells) is factored into two output matrices: the Amplitude matrix,$A \in \mathbb{R}^{n\,x\,k}$ and the Pattern matrix, $P \in \mathbb{R}^{k\,x\,m}$ [17]. Here k represents the number of latent factors or patterns learned by CoGAPS. The A matrix indicates the relative effect of each of the genes on different latent patterns while the P matrix represents the relative pattern weights for each of the cells. CoGAPS uses an atomic prior that is designed to model the non-negativity and sparsity of single cell data in learning the values for these matrices. The patterns learned by CoGAPS capture the underlying biological processes in the data[7,20] and are well suited for analyzing single cell datasets [7] [21].All analyses described in this manuscript are generated using the CoGAPS package version 3.5.8 with nIterations = 50000, sparseOptimization =TRUE, nSets = 6 (single cell parallelization)and a range of dimensionality as described in the results.

ACTIONet: Multiresolution network reconstruction analysis of single cell data and cell type identification

ACTIONet is a multiresolution matrix decomposition method that combines archetypal analysis with network reconstruction to provide a low dimensional representation of the data. ACTIONet is used for network based analysis of single cell datasets [23], [24]. In the initial step ACTIONet decomposes the input gene expression matrix into a cell influence matrix (C) and a cell state encoding matrix (H) [13]. The output matrix H, represents the relative contribution of the latent patterns towards each of the cells while the signature profile matrix indicates relative weights of patterns for each of the genes. Here S indicates the log counts of the input gene expression matrix. The decomposition is repeated for all dimensions starting from 2 up to a specified value k. The dominant patterns across all the dimensions are then collapsed to provide multi-level cell encoding matrices . In the final step ACTIONet develops a knn network based on a cell-cell similarity structure of the multilevel encoding. We apply the same procedure for pattern marker statistics used to annotate cell types from features of the CoGAPS analysis to the *W* and *H* matrices learned at each dimension with ACTIONet.

Variational Autoencoder for single cell analysis

A Variational Autoencoder (VAE) is an unsupervised machine learning method that uses a neural network to learn a reduced latent space for representing high dimensional data. VAEs have been widely used for dimensionality reduction and analysis of scRNA-seq datasets [25], [26] .The standard VAE outputs a posterior distribution which provides a

weighted summary of contribution of the latent factors towards the samples (cells). However it does not provide any interpretable link between the genes and the latent patterns, which is crucial for cell type annotations and other downstream analyses. Hence we used a linear decoded VAE (LDVAE) which uses linear functions to associate latent factors to the genes in the input data. The LDVAE is a part of the Single Cell Variational Inference (scVI) model [18] which is a probabilistic approach that implements VAEs for analyzing high dimensional single cell data.

The LDVAE uses the input gene expression matrix to train a neural network to output two matrices: the Z matrix provides a weighted summary of the contribution of each of the latent factors towards the cells and the W matrix indicates the relative weights of each of the factors towards the genes. The number of epochs, desired latent resolution and number of hidden layers are some of the important parameters that define the dimensionality of the latent space for LDVAE.

Cell type identification from the unsupervised learning models

For the analyses in this paper, it is critical to associate the low dimensional features learned through each of the dimensionality reduction techniques with discrete cell types. Analysis to distinguish cell types depends critically on associating features with specific marker genes. However, a feature of latent space methods is that they provide continuous gene signatures that explain the data when combined in linear combinations that are not readily suited to these single gene-based analyses [7]. The patternMarker

(PM) statistic is a function provided in the CoGAPS package, which associates a discrete set of genes to each of the features from the gene weights matrix in one of the matrices from the factorization [22] analogous to the D-score for NMF [27]. The statistic is used to rank genes from highest to lowest unique association with each of the patterns. For example, in the case of the A matrix from CoGAPS this statistic is given by:

$$s_{ij}\,(\mathrm{w}) \;=\; \sqrt{\frac{A_{ik}}{max\;A_i} - w_{jk}}$$

$$s = score, w \;=\; weights\;for\;patterns, A = Amplitude\;matrix$$

When adapted to the cells weights matrix, the PM statistic ranks patterns based on their unique association to the cells. This unique association is critical for determining the most relevant set of genes / cells for any associated with a feature and subsequently annotate cell types identified with CoGAPS.

We apply the pattern marker statistic through a two-step approach for cell type annotations from the factorization. First, the PM statistic applied to the matrix of gene weights matrix (A for CoGAPS, W matrix for ACTIONet, and the Z matrix for LDVAE) to rank marker genes with respect to each of learned features. If a pattern has a strong association to a set of marker genes for a given cell type , it is annotated as being associated with that cell type. In the second step, the PM statistic is applied to the matrix of cell weights (P for CoGAPS, H matrix for ACTIONet, and W for LDVAE) and cells that are strongly associated to a pattern are annotated as the cell type defined through the amplitude matrix. For instance, if the marker genes of monocyte cells are strongly

associated with the PM genes from the 5th column of the gene weights matrix then all cells that the PM statistic to 5th row of the P matrix are annotated as monocytes. As we and others have described previously, factorizations learn features can simultaneously delineate cell types, cell state transitions, and technical artifacts within a single dataset [27], [6], [7]. Therefore, features may not associate with marker genes for a single cell type. To enable cell type annotation, we filter such patterns from analysis and use the most highly associated cell type pattern to annotate every cell.

Data

The dataset used for analysis is a publicly available scRNA seq dataset of human peripheral mononuclear blood cells (PBMC) [19] . The original experiment was conducted across 92000 cells over three different sample types using different RNA-seq technologies including Drop-seq, Smart-seq2, CEL-Seq2, Seq-Well ,inDrops and 10x Chromium. We have limited our analysis to the 31021 PBMC cells across these technologies from the samples in this dataset, summarized in Supplemental Table 1.

Results

Multidimensional Analysis of CoGAPS

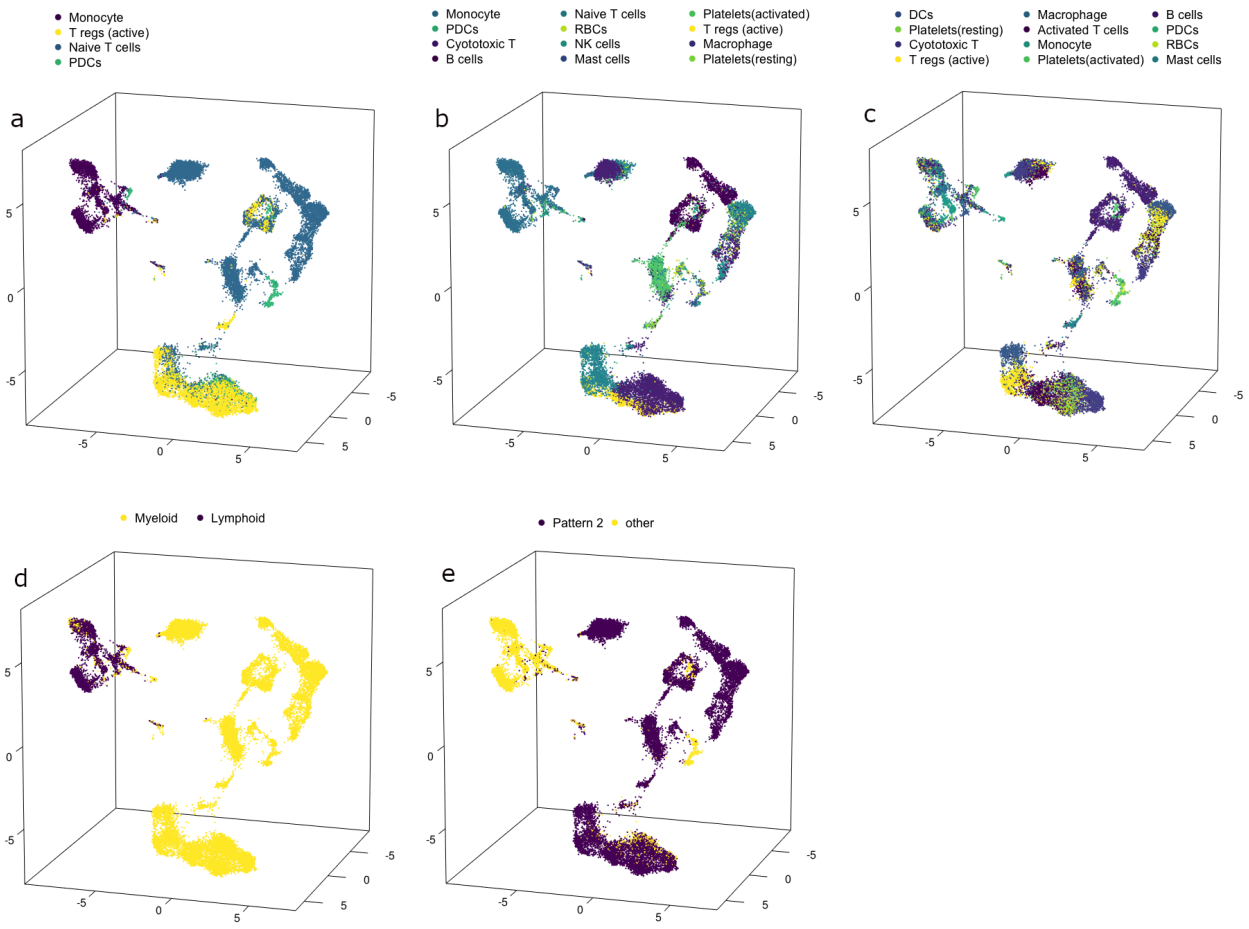

***Figure 1**: 3D UMAP representation of cell type annotations of CoGAPS results for dimensions (a) 5 , (b) 20 and (c) 55. (d) Separation of myeloid and lymphoid cell types for dimension 5 and (e) pattern 2 observed at dimension 5.*

We first sought to identify the impact of dimensionality on cell types in the reference PBMC scRNA-seq dataset, which was inferred by each of the latent space methods. We started by benchmarking this approach with CoGAPS Bayesian NMF analysis, due to the interpretable nature ensured by the factorization and described previously [6]. To analyze the impact of dimensionality, we performed CoGAPS analysis on the PBMC dataset for multiple dimensions starting from 5 up to 60 (at intervals of 5). The PM statistic was used to annotate cell types by comparing the genes uniquely associated with the learned features at each dimension to established marker genes for cell types. We observed that the number of cell types learned increased with increase in dimension. The lowest dimension expressed 4 distinct cell types (Fig 2a), which increased to 12 for dimension 20 (Fig 2b) and stayed constant up to dimension 55 .Our analysis demonstrates that cell types remain stable after dimension 40, marking this as the optimal lowest dimension with consistent cellular annotation for the PBMC dataset. A total of 12 cell types shown in (Fig 2c) including B cells, cytotoxic T cells, activated T cells, NK cells, mast cells, macrophages, monocytes, dendritic cells, Plasmacytoid Dendritic Cells(PDCs), red blood cells, platelets (resting) and platelets (activated) were observed for this dimension. At higher dimensions a larger number of patterns did not correlate to any single cell type indicative of noise or other biological processes further supporting the selection of 40 as the optimal dimension cell type identification from CoGAPS analysis of PBMCs.

While dimension 40 delineates the most discrete cell types, we observed that the cell types classified from the factorization at different dimensions match the inherent cell type hierarchies of PBMCs. At dimensions 5 and 10, the cells are broadly classified into

monocytes, naive T cells and PDCs that belong to the base of the hierarchy tree. For instance, that pattern 2 at dimension 5 (Fig 2e) differentiates myeloid from lymphoid cell types (Fig 2d). As dimensions increase, CoGAPS learns more nuanced cell types such as RBCs, cytotoxic T cells and macrophages. Thus, CoGAPS preserves the cell type hierarchies with increasing dimensions suggesting that an ensemble of factors across dimensions may more accurately reflect the biology of immune cell types in PBMCs than a single dimensionality.

Multidimensional Analysis of ACTIONet

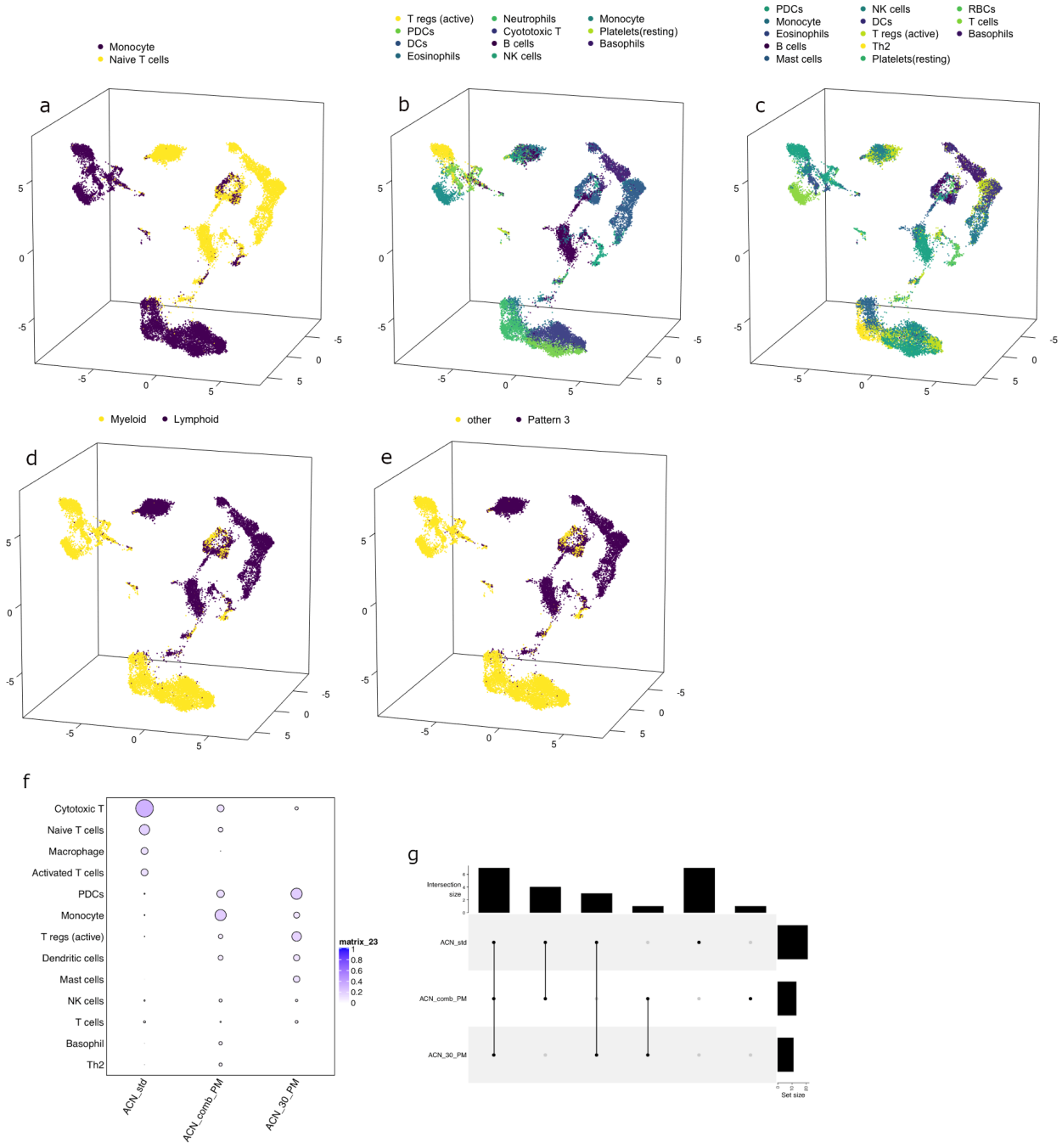

***Figure 2**:3D UMAP representation of cell type annotations of ACTIONet results for dimensions (a) 3 , (b) 15 and(c) 24. (d) Separation of myeloid and lymphoid cell types for dimension 3 and (e) pattern 2 observed at dimension 3. (f) Comparison of annotations*

To address the hierarchy of immune cell types, a recent network based dimensionality reduction algorithm ACTIONet was developed to implicitly incorporate factorizations from multiple dimensions [13]. To explore the impact of the multiple dimensional resolution from this method, we next performed an analysis with this method to evaluate the sensitivity of cell type calls obtained at each dimension. Specifically, this ACTIONet analysis was conducted on the same PBMC dataset for varying dimensions ranging from 2 to 30. While CoGAPS limits pattern weights to non-negative values, ACTIONet allows for negative weights as well. Therefore, we hypothesized that half the number of features in ACTIONet would give an equivalent representation of the space to CoGAPS. This guided the selection of 30 dimensions as the maximum for the ACTIONet analysis, as it represents half of the maximum 60 dimensions analyzed in CoGAPS.

To enable direct comparison between CoGAPS and ACTIONet, we adapted the CoGAPS PM statistic to cell type classification from ACTIONet. We observed that the number of cell types learned similarly increased with dimension for ACTIONet. However, relatively lower number of cell types were resolved below dimension 10. As an example, we illustrate that only 2 cell types were observed at the 3 dimensional factorization with ACTIONet (Fig 3a). The cell type classifications are inconsistent in factorizations from dimensions between 3 and 10. The total number of cell types learned peaks at a dimension of 24 (Fig 3c) and the cell types learned then remain constant for higher dimensions. Thus, we called 24 the optimal dimension for annotating cell types in the PBMC dataset. The cell types learned at this dimension B cells, T cells, Tregs (active), Th2, NK cells, monocytes, eosinophils, basophils, mast cells ,Plasmacytoid Dendritic

Cells (PDCs), dendritic cells, red blood cells and platelets (resting). While specific cellular subtypes varied at low dimensions, we observed that lower dimensions of the ACTIONet analysis also separates myeloid and lymphoid cell types. Notably, Figs 3d and 3e show that pattern 3 for dimension 3 differentiates myeloid cell types from lymphoid cells. Thus, while specific cell type annotations are less consistent than CoGAPS, this analysis suggests that ACTIONet is also identifying a similar hierarchy of cell types in factorizations across multiple dimensions.

Comparison with ACTIONet annotation function

We compared the cell type annotations obtained from PM statistic with ACTIONet's in-built function for annotating using marker genes. ACTIONet annotates the data by collapsing dominant patterns learned across all dimensions to form a consolidated cell encoding matrix which is used to annotate the cells. We compared the annotations with annotations using our own method for the optimal dimension of 24 as well as the PM statistic based annotations of the collapsed cell encoding matrix.The upset plot in fig 3g shows that both the PM based annotation sets have a fairly large intersection set with the ACTIONet method, thus indicating the annotations are comparable. However we observed that the total number of cell types learned and the proportion of each cell type varies between ACTIONet and PM based methods (fig 3f). ACTIONet learns more number of cell types (21) when compared to the PM based annotations (12 or 13). Furthermore , cytotoxic T cells are highly expressed only in ACTIONet's method while monocytes and PDCs are highly expressed by PM.

Multidimensional Analysis of LDVAE

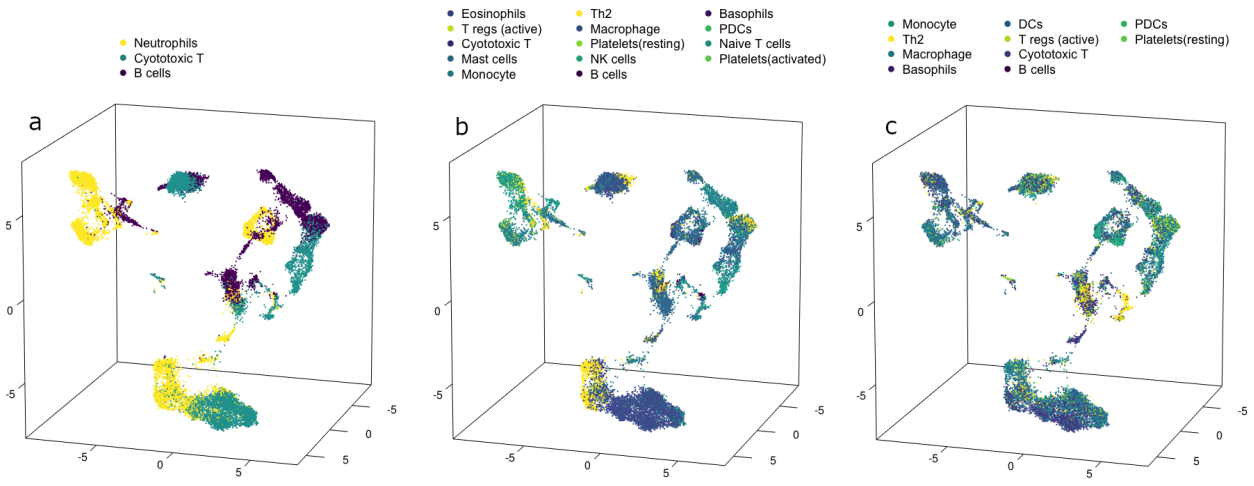

Figure 3:3D UMAP representation of cell type annotations of CoGAPS results for dimensions (a) 5 , (b) 15 and (c) 30.

.

Additional, autoencoders are emerging as a critical class of dimension reduction algorithms for single cell data. Therefore, to further compare the role of these non-linear decompositions we also applied a VAE model to the same PBMC data. We used a variant with a linear decoding step, LDVAE [18]. to enable direct comparison of this approach with the cell type annotations from the PM statistic in the linear factorizations from CoGAPS and ACTIONet. Specifically, analyzed the LDVAE model for varying dimensions from 5 up to 30. As we described for ACTIONet, the maximum dimension of 30 was selected to be comparable with the 60 dimensions from CoGAPS as to account for the negative weights of the factorization from that model. As in the previous two methods, the number of cell types learned increased with dimension. In LDVAE, number

of cell types learned peaked at a dimension of 25. However, the annotations varied between dimensions and did not correspond to the cell type hierarchies in immune cells or differentiate the myeloid and lymphoid cell types.

Comparison of Cell Types Across Methods at Select Dimensions

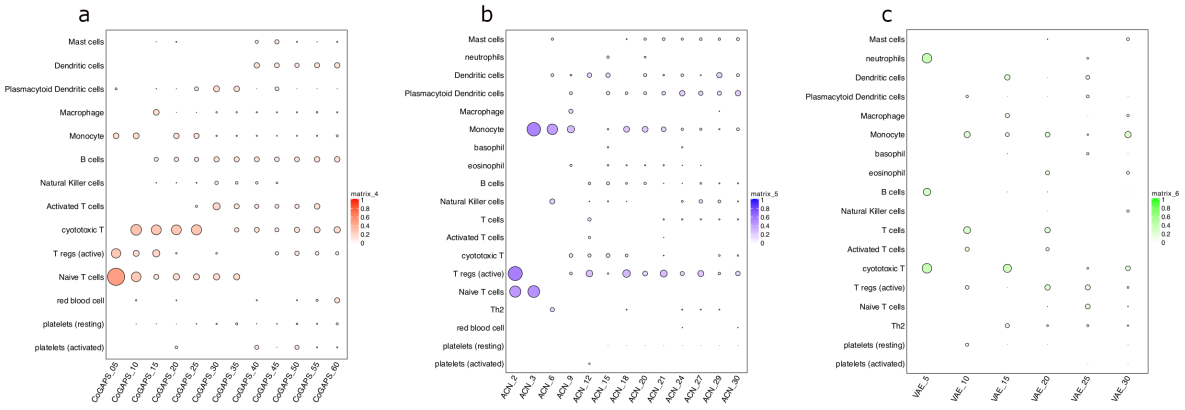

Figure 4:Comparison of proportion of different cell types expressed across dimensions for (a)CoGAPS, (b) ACTIONet and (c)LDVAE.

If all methods are equally identifying features associated with the biological properties of immune cells, they should be obtaining similar classification of cell types. Notably, we observed that all three methods followed the general trend of expressing more cell types with increasing dimensions. Notably, at lower dimensions both CoGAPS and ACTIONet delineate myeloid and lymphoid cells and then all methods delineate more discrete immune cell types at higher dimensions. The number of cell types learned peaks at a certain dimension and remains stable for subsequently higher dimensions. We establish this dimension as the optimal cell dimension for each method, and it is approximately

twice as high in CoGAPS which allows for only non-negative feature weights (40, Fig 5a) as ACTIONet (24, Fig 5b) or LDVAE (25, Fig 5c) which allow for both positive and negative feature weights.

The total number of distinct cell types expressed across all dimensions was highest for ACTIONet (19 cell types), while CoGAPS expressed a total of 14 cell types and LDVAE expressed 17 cell types. In addition to the number, the specific cell types identified with the PM statistic also varied across the methods. We found a significant difference in the percentage of myeloid and lymphoid cell types expressed by CoGAPS and ACTIONet. Overall, CoGAPS expressed a higher percentage of lymphoid cell types such as B cells and naive T cells while ACTIONet expressed a higher percentage of myeloid cell types such as monocytes and mast cells.Some of the myeloid cell types such as eosinophil, neutrophil and basophil were expressed only by ACTIONet.The difference was especially prominent especially at lower dimensions. While ACTIONet expressed 53.14%, 56.12% 74% of myeloid cell types at dimensions 3, 5 and 9 respectively, CoGAPS expressed 22.8% ,17.32% and 19.72% dimensions 3, 5 and 9 respectively. The VAE cell annotations did not express any specific trend of differences in percentages of myeloid and lymphoid cell types. The percentage of both cell types varied greatly across dimensions.

Consistency of Cell Type Annotations across Dimensions

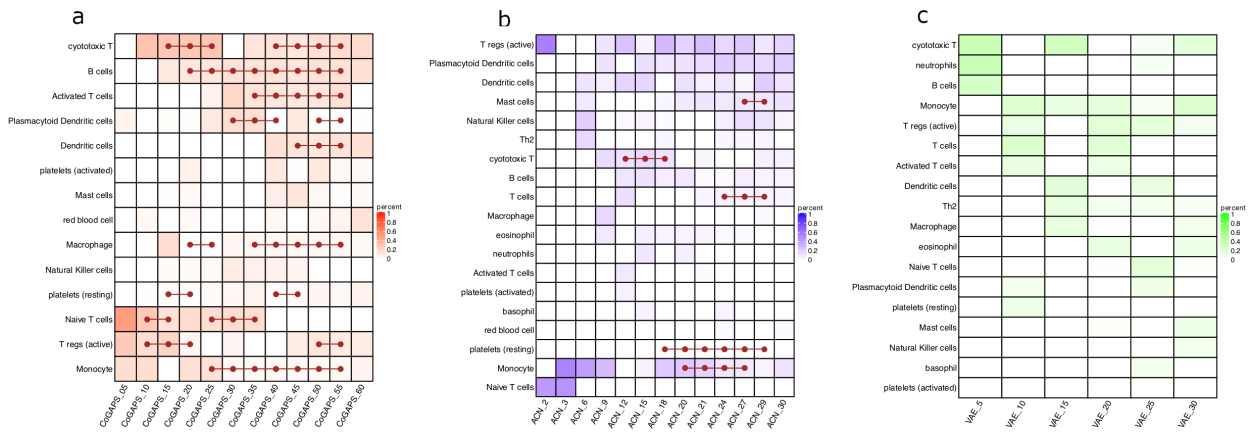

Figure 5:Comparison of consistency cell types annotations across dimensions for (a)CoGAPS, (b) ACTIONet and (c)LDVAE.

We observed that the annotations of cells varied from one dimension to another within each of the methods, with stable separation of myeloid and lymphoid cells at low dimensions and then stable annotations of additional cell types at higher dimensions for both CoGAPS and ACTIONet. To quantify this similarity between dimensions, we calculation the proportion of cell types with consistent annotations between consecutive dimensions and call the annotations consistent if at least 75% of the cells annotated at a lower dimension match the next higher dimension. Overall, cell type calls from CoGAPS were consistent across dimensions. The consistency generally increases with dimension as the new cell types are learned with increasing dimensions. However the cell type annotations become stable for higher dimensions after dimension 35 (Fig 6a). Notably, the annotations of B cells, monocytes, macrophages, cytotoxic T and activated T cells are consistent for dimensions 25 through 55. The annotations for mast cells, red blood cells and platelets (activated) were not consistent even at higher dimensions. We observed that ACTIONet had a lower consistency in cell type annotations across

dimensions. At a threshold of 0.75, only monocytes and platelets (resting) remained consistent for dimensions 18 through 27 (Fig 6b). Decreasing the threshold to 0.5 resulted in consistent annotations of monocytes, platelets (resting) and cytotoxic T cells for dimensions 15 to 27. We could not find any significant consistency between cell type annotations across any dimensions for LDVAE at a threshold of 0.75 (Fig 6c). LDVAE expressed consistency only for monocytes for dimensions 10 to 25 even at very low thresholds of 0.10.

Cell Type Hierarchies across Dimensions

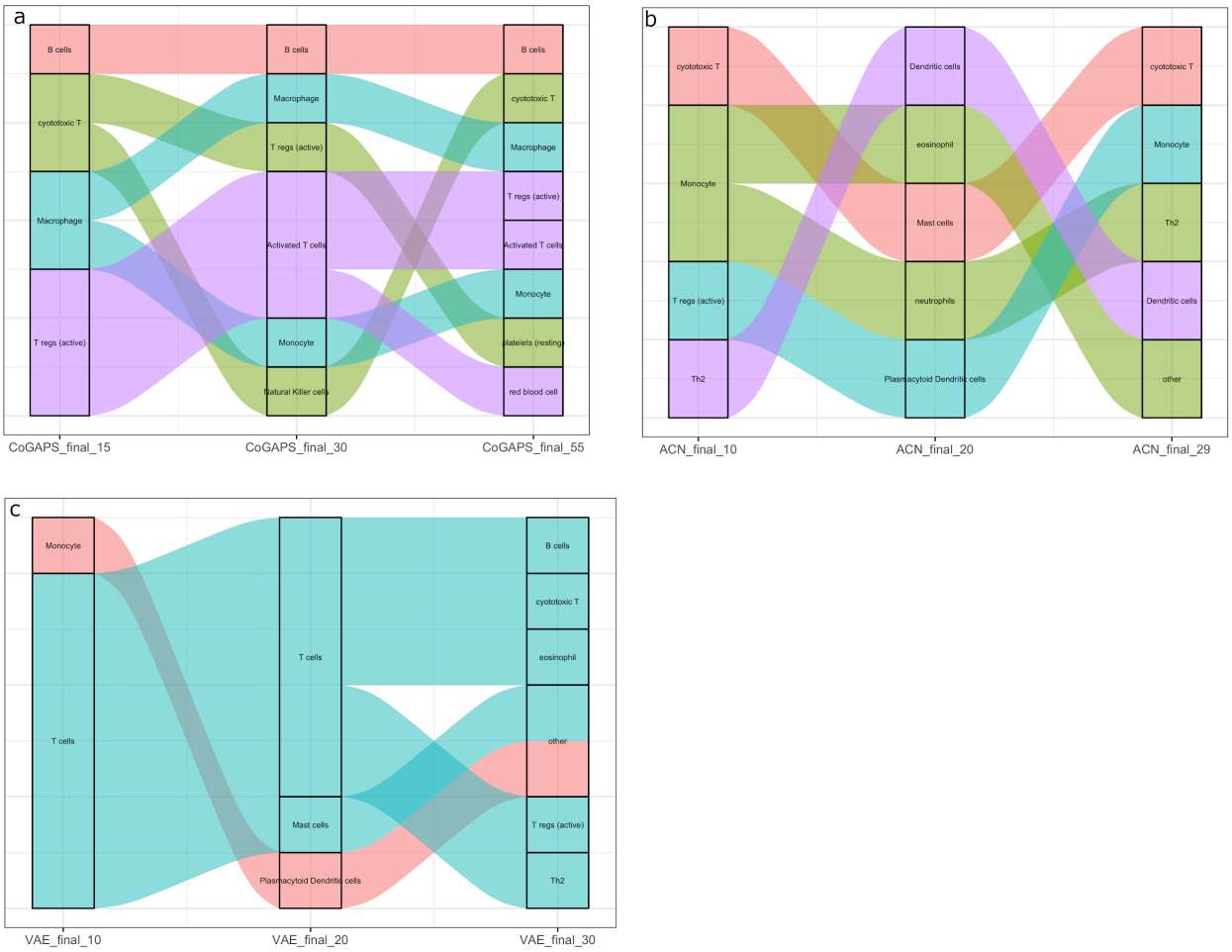

Figure 6:Comparison of cell types hierarchies across dimensions for (a)CoGAPS, (b) ACTIONet and (c)LDVAE.

An important aspect of the PBMC dataset is the cell type hierarchies that are observed within the myeloid and lymphoid cell types. We investigated whether these hierarchies are preserved in each of the three dimensionality reduction methods.

We observed that CoGAPS preserves the cell-type hierarchies very well. We studied the branching structure of the cell types from dimensions 15 to 30 and 30 to 55 (Fig 7a). Some of the prominent hierarchies included monocytes branching out into macrophages, T regs (active) splitting into activated T cells and RBCs and cytotoxic T cells branching into NK cells and platelets (resting). These branching structures are in accordance with the known cell type hierarchies of PBMCs.

ACTIONet preserved some of the underlying cell type hierarchies. When navigating from dimension 10 to 20 and 20 to 29 (fig 7b), it was observed that cytotoxic T cells branch into mast cells.However, monocytes which are expected to branch into macrophages instead split into Th2 and neutrophils. We also noticed that Tregs (active) branch into monocytes, which deviates with current understanding of immune cell lineages in PBMCs. Lastly, the LDVAE results across dimensions expressed less consistency in cell types and hence did not show any prominent branching structure. Fig 7c shows how T cells branch into numerous cell types including eosinophil and B cells, which are not usually observed in PBMCs.

Discussion

We have compared the effect of dimensionality and factorization methods on the latent space obtained and corresponding biological annotation of single cell data from PBMCs using three prevalent dimensionality reduction methods: CoGAPS [17], ACTIONet [13], and LDVAE [18].We observed that resolution of annotated cell type increased with dimensionality for CoGAPS and ACTIONet. Notably, the number of cell types identified

peaked at a particular dimension and remained consistent for higher dimensions providing a metric to assess optimal dimensionality. However, the cell types from LDVAE did not demonstrate similar consistency across the dimensions. While there was an optimal dimension for identifying high resolution cell types, we observed that both ACTIONet and CoGAPS identified lower resolution cell types that preserve established immune lineages. Most notably, the lymphoid branch of CoGAPS analysis diverged into T cells and then T cell subtypes as the dimensionality increased over a range. This observation suggests that while a single dimension can be established to optimize the resolution of cell type classification, no single dimensionality is necessarily ideal for capturing all the cellular features from a single cell dataset with a latent space method consistent with observations in previous studies in bulk RNA-seq datasets [9]. This observation suggests that the multi-resolution approach for cell type classification developed in network based methods such as ACTIONet [13] is widely applicable to latent space methods, including notably NMF methods such as CoGAPS [17].Identification and annotation of cell types is an integral part of single cell analysis, to which low dimensional latent space methods are a critical component. Most commonly, these factorization methods are used to provide interpretable cell groupings for clustering from which marker genes are then assigned. However, a component of these analyses is their corresponding gene weights, which if interpretable can also be used for functional annotation of learned features [7]. In the case of cell type classification, this relies on association of the learned gene weights with established markers of cell types. We have proposed the pattern marker statistic as an efficient measure to annotate cell types by combining the effect of the latent factors on marker genes as well as the samples. Whereas many gene inference

methods from latent space methods rely on the relative magnitude of weights of genes in each feature, the pattern marker statistic instead ranks genes based on their unique association with a feature [22]. By relying on uniqueness, this statistic allows for factorization methods that model shared gene expression between multiple biological processes and filter highly expressed genes that may bias all factors [7].Similar gene rankings on uniqueness have been shown to be critical to identify specific marker genes for latent spaces that represent the heterogeneity in single cell datasets [27].A number of advanced annotation methods based on supervised learning [28], [29] and probabilistic gene expression [30] often prove to be computationally intensive and require knowledge of a reference database with identified cell types. PM statistic requires no reference dataset and enables direct comparison to established marker genes. Future analysis could be extended to comparing the efficiency of PM statistic to other marker gene annotation techniques [31] or comparison with additional methods for cell type identification [32]. Notably, we anticipate that while the PM based marker gene approach is well suited to the identification of established immune cell types, the gene weights in the entire gene signature are critical to identify biological processes associated with cell state transitions or more rare cellular subpopulations without well-established marker genes.

We note that the PBMC dataset [19] selected in this paper was selected due to the well-established lineage and cellular types upon which to benchmark the performance of the disparate latent space methods and dimensionalities. A further unique aspect of this dataset is the inclusion of the same cell types with multiple measurement technologies

and single cell library preparations. We note the robustness of the cell type classifications across dimensionalities in spite of the technical artifacts suggests that both CoGAPS and ACTIONet identify features associated with biological processes such as cell types independently of the technical artifacts. It is critical in future work to evaluate the role these technical covariates play on the features learned by each of these methods and whether they contribute to the low stability of cellular annotations across dimensions observed in LDVAE. These covariates may also contribute differently to the features learned from different methods, which may contribute to the observed discrepancies in cellular annotations between CoGAPS and ACTIONet and performance biases in cell type classification for specific technologies between the methods.

An asset of the cross-method comparison in this study is the observation of distinct cellular populations from distinct methods, even at the optimal dimensionality established for maximal cell type resolution in each method. Notably, CoGAPS resolved more myeloid cell types and ACTIONet more lymphoid cell types. Many benchmark studies use this discrepancy to rank methods to select a single, optimal method for cell type classification [32]. However, we hypothesize that the difference between the cell types resolved by the different methods may result in optimal classification using an ensemble approach which combines significant patterns across different methods as well as the multi-resolution from multiple dimensions.Over the recent years a number of ensemble approaches combining clustering methods [33] random projection [33] and machine learning techniques [34] have been developed for single cell analysis.Future work combining

patterns across significant patterns across different methods could better resolve the distinct biological features in the data hidden through a single factorization.

Bibliography

[1] Zhang Y-Q and Rajapakse J C 2008 Machine learning in bioinformatics (Hoboken, NJ, USA: John Wiley & Sons, Inc.)

[2] Becht E, McInnes L, Healy J, Dutertre C-A, Kwok I W H, Ng L G, Ginhoux F and Newell E W 2018 Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol.

[3] Moon K R, van Dijk D, Wang Z, Gigante S, Burkhardt D B, Chen W S, Yim K, Elzen A van den, Hirn M J, Coifman R R, Ivanova N B, Wolf G and Krishnaswamy S 2019 Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. **37** 1482–92

[4] Jain A K, Duin P W and Jianchang Mao 2000 Statistical pattern recognition: a review IEEE Trans. Pattern Anal. Mach. Intell. **22** 4–37

[5] Cleary B, Cong L, Cheung A, Lander E S and Regev A 2017 Efficient generation of transcriptomic profiles by random composite measurements. Cell **171** 1424-1436.e18

[6] Stein-O'Brien G L, Clark B S, Sherman T, Zibetti C, Hu Q, Sealfon R, Liu S, Qian J, Colantuoni C, Blackshaw S, Goff L A and Fertig E J 2019 Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues,

and Species. Cell Syst. **8** 395-411.e8

[7] Stein-O'Brien G L, Arora R, Culhane A C, Favorov A V, Garmire L X, Greene C S, Goff L A, Li Y, Ngom A, Ochs M F, Xu Y and Fertig E J 2018 Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. **34** 790–805

[8] van LVDMAATENGMAIL.COM L Visualizing Data using t-SNE

[9] Way G P, Zietz M, Himmelstein D S and Greene C S 2019 Sequential compression across latent space dimensions enhances gene expression signatures BioRxiv

[10] Bouveyron C, Latouche P and Mattei P 2020 Exact dimensionality selection for Bayesian PCA Scand. J. Stat. **47** 196–211

[11] Kelton C J, Lee W, Rusay M, Maxian O, Fertig E J and Ochs M F 2015 The estimation of dimensionality in gene expression data using Nonnegative Matrix Factorization 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE) pp 1642–9

[12] Yeung K Y and Ruzzo W L 2001 Principal component analysis for clustering gene expression data. Bioinformatics **17** 763–74

[13] Mohammadi S, Davila-Velderrain J and Kellis M 2019 Multi-resolution single-cell state characterization via joint archetypal/network analysis BioRxiv

[14] Borges H B and Nievola J C 2012 Comparing the dimensionality reduction methods in gene expression databases Expert Syst. Appl. **39** 10780–95

[15] Bartenhagen C, Klein H-U, Ruckert C, Jiang X and Dugas M 2010 Comparative

study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC Bioinformatics **11** 567

[16] Sun S, Zhu J, Ma Y and Zhou X 2019 Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. **20** 269

[17] Fertig E J, Ding J, Favorov A V, Parmigiani G and Ochs M F 2010 CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics **26** 2792–3

[18] Lopez R, Regier J, Cole M B, Jordan M I and Yosef N 2018 Deep generative modeling for single-cell transcriptomics. Nat. Methods **15** 1053–8

[19] Ding J, Adiconis X, Simmons S K, Kowalczyk M S, Hession C C, Marjanovic N D, Hughes T K, Wadsworth M H, Burks T, Nguyen L T, Kwon J Y H, Barak B, Ge W, Kedaigle A J, Carroll S, Li S, Hacohen N, Rozenblatt-Rosen O, Shalek A K, Villani A-C and Levin J Z 2019 Systematic comparative analysis of single cell RNA-sequencing methods BioRxiv

[20] Sibisi S and Skilling J 1997 Prior distributions on measure space J. Royal Statistical Soc. B **59** 217–35

[21] Erbe R, Kess;ler M D, Favorov A V, Easwaran H, Gaykalova D A and Fertig E J 2020 Matrix factorization and transfer learning uncover regulatory biology across multiple single-cell ATAC-seq data sets BioRxiv

[22] Stein-O'Brien G L, Carey J L, Lee W S, Considine M, Favorov A V, Flam E, Guo T, Li S, Marchionni L, Sherman T, Sivy S, Gaykalova D A, McKay R D, Ochs M F,

Colantuoni C and Fertig E J 2017 PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. Bioinformatics **33** 1892–4

[23] Menon M, Mohammadi S, Davila-Velderrain J, Goods B A, Cadwell T D, Xing Y, Stemmer-Rachamimov A, Shalek A K, Love J C, Kellis M and Hafler B P 2019 Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. Nat. Commun. **10** 4902

[24] Mohammadi S, Davila-Velderrain J and Kellis M 2019 Reconstruction of Cell-type-Specific Interactomes at Single-Cell Resolution. Cell Syst. **9** 559-568.e4

[25] Wang D and Gu J 2017 VASC: dimension reduction and visualization of single cell RNA sequencing data by deep variational autoencoder BioRxiv

[26] Grønbech C H, Vording M F, Timshel P N, Sønderby C K, Pers T H and Winther O 2018 scVAE: Variational auto-encoders for single-cell gene expression data BioRxiv

[27] Zhu X, Ching T, Pan X, Weissman S M and Garmire L 2017 Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. PeerJ **5** e2888

[28] Pliner H A, Shendure J and Trapnell C 2019 Supervised classification enables rapid annotation of cell atlases. Nat. Methods **16** 983–6

[29] Alquicira-Hernandez J, Sathe A, Ji H P, Nguyen Q and Powell J E 2019 scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. **20** 264

[30] Grabski I N and Irizarry R A 2020 Probabilistic gene expression signatures identify

cell-types from single cell RNA-seq data BioRxiv

[31] Zhang A W, O'Flanagan C, Chavez E A, Lim J L P, Ceglia N, McPherson A, Wiens M, Walters P, Chan T, Hewitson B, Lai D, Mottok A, Sarkozy C, Chong L, Aoki T, Wang X, Weng A P, McAlpine J N, Aparicio S, Steidl C and Shah S P 2019 Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat. Methods **16** 1007–15

[32] Huang Q, Liu Y, Du Y and Garmire L 2019 Evaluation of Cell Type Deconvolution R Packages on Single Cell RNA-seq Data BioRxiv

[33] Yang Y, Huh R, Culpepper H W, Lin Y, Love M I and Li Y 2019 SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data. Bioinformatics **35** 1269–77

[34] Chen X, Chen S and Jiang R 2019 EnClaSC: A novel ensemble approach for accurate and robust cell-type classification of single-cell transcriptomes BioRxiv

Curriculum Vitae

Archana Balan (email: abalan2@jh.edu)

Education

MSE Biomedical Engineering, Johns Hopkins University (Aug 2018 – May 2020)
Focus Area: Computational Medicine
Dissertation Title: Multi-resolution analysis and convergence testing of CoGAPS algorithm.

B.Tech Electronics and Instrumentation Engineering, VIT University, India (Jul 2018- May 2017)
Second rank holder (in a class of 78 students)
Capstone Project Title: Wearable Sensor for ECG Measurements

Research Experience

Graduate Researcher, Johns Hopkins University (Jul 2019 – Jul 2020)
Principal Investigator: Dr.Elana Fertig (Fertig Lab)

- Effect of higher dimensionalities on biological relevance of features learned in the NMF method.
- Analysis on multiple datasets including PBMC, brain and mouse retina datasets
- Exploring GPU computing for Bayesian NMF algorithms.

Project Assistant, Indian Institute of Sciences(IISc), India (Jul 2017 – Jul 2018)
Principal Investigator: Dr.Radhakant Padhi (Aerospace Department, IISc)

- Development of Artificial Pancreas for the Indian population.
- Quantitative modeling and parameter estimation for glucose-insulin regulation system.
- Project was awarded **Chellaram Diabetes Foundation Innovation in Diabetes Research Award 2018**
-

Publications

1. Biradar, S., Balan, A., Padhi, R. and Dharamalingam, M., 2019. Modified Bergman minimal model for glucose-insulin dynamics and estimation of model parameters for Indian population. DiabetesTechnology &Therapeutics, 21(1), p.A80.
2. Nath, A., Biradar, S., Balan, A., Dey, R. and Padhi, R., 2018. Physiological models and control for type 1 diabetes mellitus: a brief review. IFAC-PapersOnLine, 51(1), pp.289-294.

3. Ramasamy, S. and Balan, A., 2018. Wearable sensors for ECG measurement: A review. Sensor Review, 38(4),pp.412-419